



Aprendizaje automático y eventos políticos

Aplicación de un enfoque semisupervisado para producir un conjunto de datos sobre gabinetes latinoamericanos

Bastián González-Bustamante

University of Oxford

✉ bastian.gonzalezbustamante@politics.ox.ac.uk

Presentación preparada para el segundo taller del laboratorio de datos

[Training Data Lab](#), 18 de noviembre de 2022

Tabla de contenidos

1. Introducción
2. Recolección y creación del conjunto de datos
3. Aprendizaje automático para clasificar las llamadas a dimisión
4. Validación y aplicaciones
5. Discusión



Artwork por DALL·E, inspirado por Mattise

Introducción

Minería de datos y aprendizaje automático

Basados en Procesamiento del Lenguaje Natural (**Natural Language Processing, NLP**) y el enfoque **text-as-data**, presentamos un conjunto de datos totalmente novedoso construido aplicando técnicas de minería de datos y aprendizaje automático.

El conjunto de datos contiene información detallada sobre **gabinetes en 12 democracias latinoamericanas desde mediados de la década de 1970 hasta la fecha**.

Utilizamos archivos de prensa digitalizados con algoritmos de extracción mediante reconocimiento óptico (**Optical Character Recognition, OCR**). Luego, con aprendizaje automático, identificamos las **peticiones a dimisiones ministeriales** durante el periodo.

Peticiones a dimisiones de ministros

Aunque la literatura ha identificado que las protestas, crisis económicas, escándalos de diversa índole, entre otros **eventos estocásticos**, afectan a la estabilidad ([Camerlo y Pérez-Liñán, 2015](#); [Martínez-Gallardo, 2014](#)), medir todos estos posibles *random shocks* es complejo y empíricamente costoso ([Berlinski et al., 2010](#)).

Por lo tanto, señales como las **peticiones a dimisión de ministros** pueden ser indicadores empíricamente eficientes. Este indicador es totalmente inédito en los sistemas de los sistemas presidenciales latinoamericanos y ofrece una serie de aplicaciones potenciales más allá del estudio de las de las coaliciones y la rotación ministerial, abriendo nuevas vías para futuras investigaciones sobre opinión pública, polarización política o políticas públicas.

Recolección y creación del conjunto de datos

Primer conjunto de datos brutos

La recolección de datos se basa en un procedimiento de tres etapas. Primero, se recopila un **conjunto de datos brutos de ministros y ministerios** en función de una serie de variables relevantes (*i.e.*, país, nombre del ministro, sexo, afiliación a un partido, fecha de nombramiento, etc.).

Este conjunto de datos brutos se recopila a partir de una revisión de fuentes oficiales, prensa reconocida, información disponible en las bibliotecas del Congreso de cada país y datos de **Latin American Weekly Report (LAWR)**.

Estos informes semanales de entre 12 y 16 páginas presentan un compendio de noticias relevantes sobre cada país de la región. En promedio, constituyen aproximadamente 600 páginas por año hasta 2002 y 800 páginas desde entonces.

Segundo conjunto de datos brutos

Segundo, se genera un **nuevo conjunto de datos brutos de llamadas semanales** a dimisión ministerial a partir de una revisión de los archivos de LAWR. Se identifican las menciones específicas de los miembros del gabinete, lo que nos permite codificar la variable Llamadas a dimisión. Para esto, los archivos de LAWR se recopilaron en tres lotes de datos diferentes.

Estos archivos se almacenan en un repositorio privado en  GitHub con acceso controlado por  2FA. Además, se realiza una copia de seguridad en  Hierarchical File Server (HFS) para su recuperación en caso de incidentes en el *hub* de la Universidad de Oxford conectado con Code42 Cloud Backup, que permite el control de versiones, restauración, copias de seguridad programadas y 256-bit Advanced Encryption Standard (AES).

Lotes de archivos de LAWR

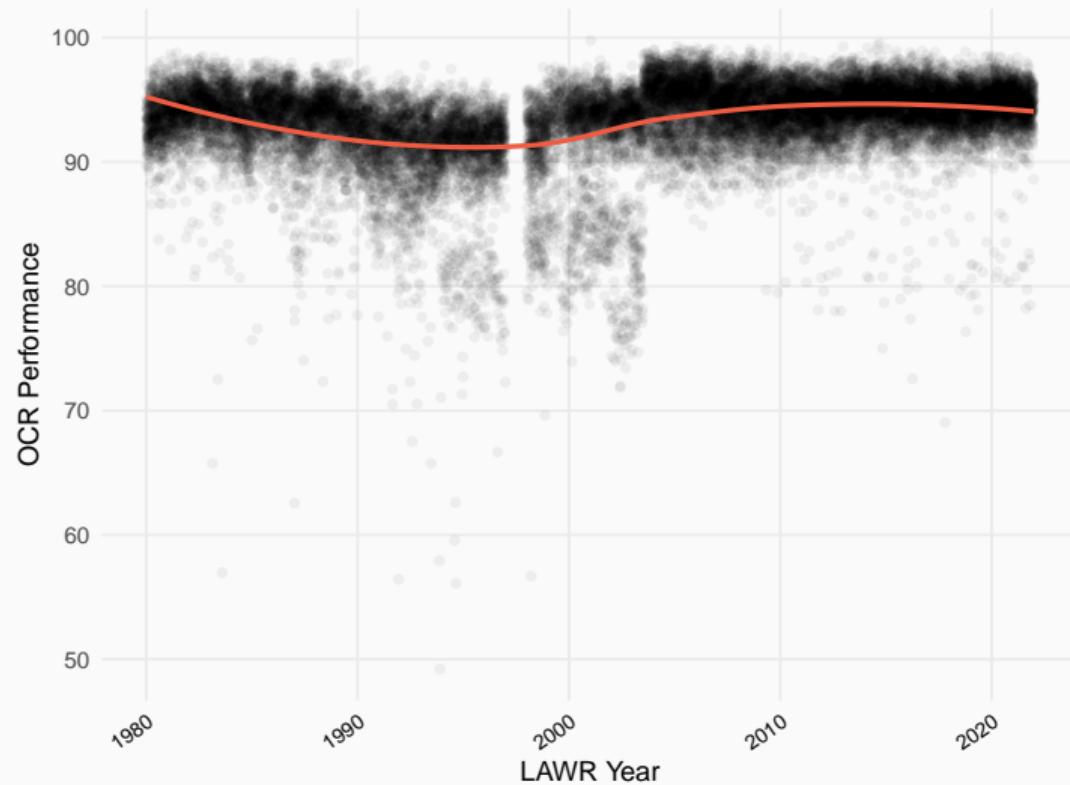
Lote	Archivos	Años	Formato	Tamaño
Lote-01	LAWR-1975 a LAWR-1979 Incluyendo 1997	5	CSV UTF-8	9,56 MB
Lote-02	LAWR-1980 a LAWR-1998 Excluyendo 1997 Incluyendo primera mitad de 2003	18,5	Imágenes PNG	38,5 GB
Lote-03	LAWR-1999 a LAWR-2021 Excluyendo primer mitad de 2003	22,5	Archivos PDF	341 MB

Nota: Recolección y revisión de 12 países latinoamericanos gracias al Fondo Muriel Wise de St Hilda's College, las suscripciones de Bodleian Libraries en Oxford y Society for Latin American Studies (SLAS) Postgraduate and Postdoctoral Research Award 2022.

Algoritmos de binarización

Nota: Las imágenes corresponden a una página de LAWR antes de la binarización y después de la aplicación de las técnicas de [Sauvola y Pietikäinen \(2000\)](#), [Singh et al. \(2011\)](#) y [Su et al. \(2010\)](#).

Precisión del algoritmo OCR



Estructura de los datos

- ➊ Aplicación OCR sobre **41 años** de archivos con un total de **28.090 páginas**.
- ➋ Identificamos **18,198 menciones** de miembros de gabinetes (incorporando el primer lote se incrementa a **19.925** y la cobertura a **46 años**).

Elaboramos un conjunto de datos semanales con llamadas a dimisión para fusionarlos con los primeros datos y otras fuentes como aprobación presidencial, variables de los sistemas de partidos e indicadores macroeconómicos.

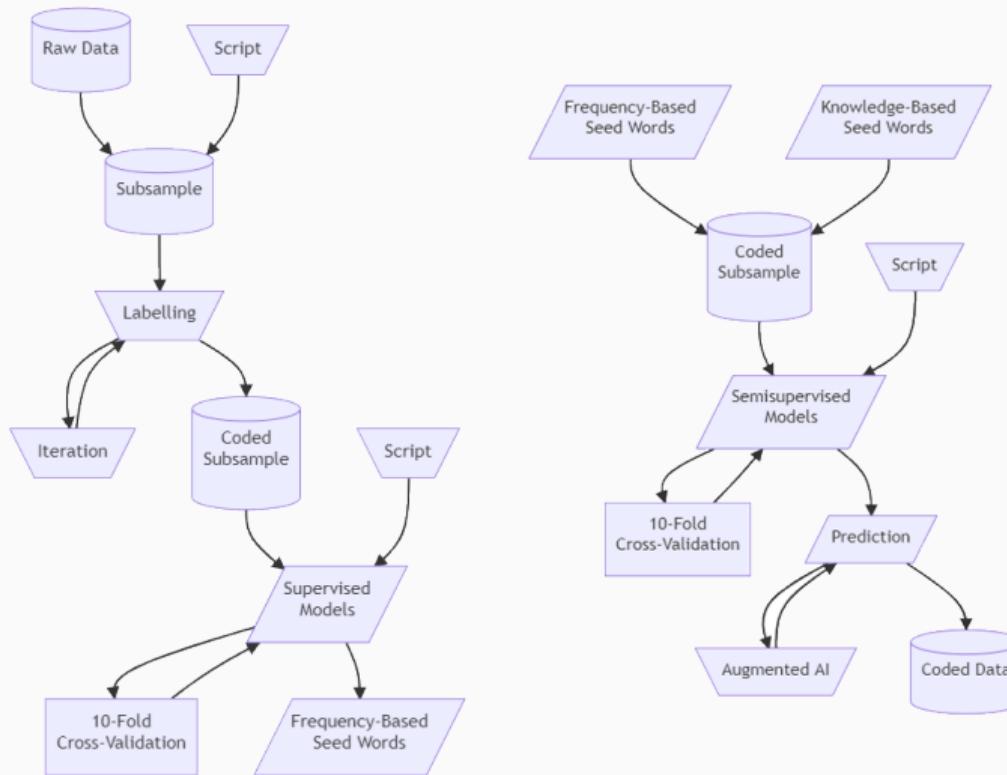
Por lo tanto, codificamos los datos como **time-dependent** dando cuenta de los cambios en el tiempo.

- ➌ Para mayor información técnica ver mi [presentación en userR conference](#).

¿Cómo distinguimos entre menciones y llamadas?

Aprendizaje automático para clasificar las llamadas a dimisión

Pipeline de los algoritmos de aprendizaje automático



Etiquetado para el entrenamiento de modelos supervisados

- Muestra aleatoria de **1.000 reportes de LAWR** considerando las menciones a miembros del gabinete. Grupo de 3.000 observaciones (tres etiquetas codificadas manualmente por reporte).
- Utilizamos la plataforma  [Labelbox \(2021\)](#), la que luego nos permitió exportar la información en JavaScript Object Notation (JSON).

El etiquetado fue realizado por **seis codificadores humanos**, lo que involucró 41 horas y 55 minutos de trabajo en la plataforma utilizando las categorías de [Dewan y Dowding \(2005\)](#): escándalos personales, financieros, fracasos de políticas públicas, desacuerdos internos u otras controversias.

Rendimiento del proceso de etiquetado

Codificador	Etiquetas	$M(T)$	Etiqueta	\sum Tiempo	Consenso
1	966		39s	10h 25m	94,05
2	718		1m 15s	15h 0m	94,84
3	704		38s	7h 29m	94,82
4	318		1m 4s	5h 38m	95,47
5	263		17s	1h 14m	99,26
6	101		1m 17s	2h 9m	88,50

Nota: Krippendorff's α (bootstrap con 1,000 iteraciones) = 0,870 y $CI_{95\%}$: 0,844 a 0,900. Considerando una clasificación binaria α = 0,895 y $CI_{95\%}$: 0,868 a 0,922.

Rendimiento de los algoritmos clasificadores

Modelos supervisados	Snowball 500 Tokens		
	Exactitud	Precisión	Recall
Naive Bayes	0,714	0,803	0,825
Kernel Linear SVM	0,658	0,802	0,728
Kernel Gaussian RBF SVM	0,756	0,756	0,999
Random Forest (100 árboles)	0,761	0,760	0,998
Random Forest (500 árboles)	0,761	0,760	0,999
XGBoost	0,784	0,811	0,932

Modelos supervisados	Stem. SMART 100 Tokens		
	Exactitud	Precisión	Recall
Naive Bayes	0,725	0,786	0,876
Kernel Linear SVM	0,735	0,785	0,894
Kernel Gaussian RBF SVM	0,766	0,766	0,993
Random Forest (100 árboles)	0,775	0,773	0,995
Random Forest (500 árboles)	0,771	0,770	0,995
XGBoost	0,771	0,809	0,912

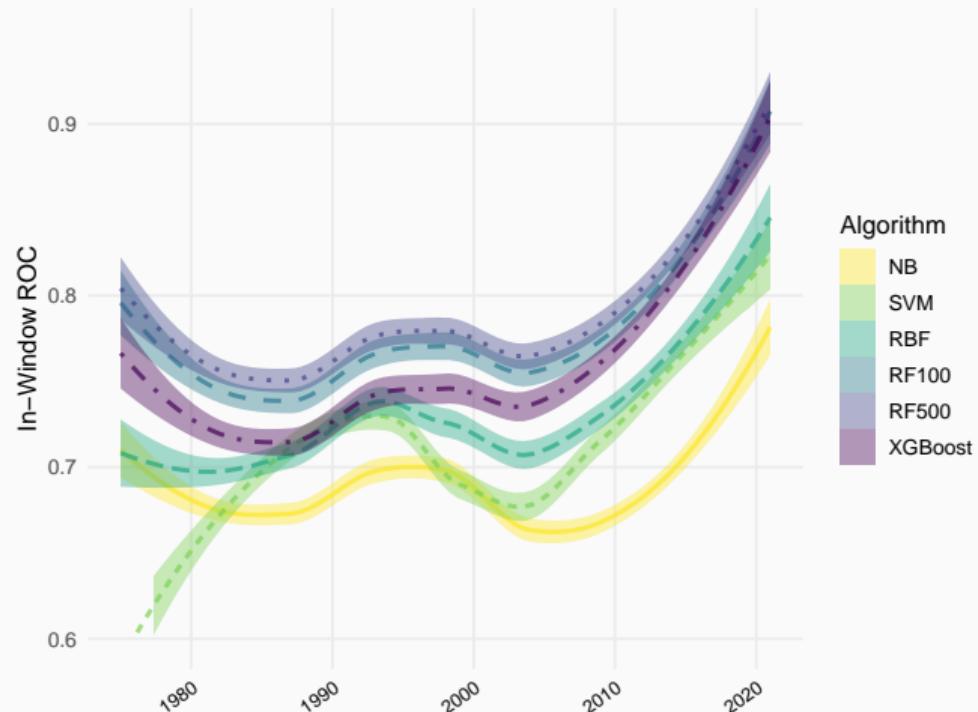
Modelos semisupervisados	Knowledge-Based Seeds		
	Exactitud	Precisión	Recall
Naive Bayes	0,713	0,734	0,948
Kernel Linear SVM	0,698	0,741	0,900
Kernel Gaussian RBF SVM	0,726	0,728	0,995
Random Forest (100 árboles)	0,743	0,742	0,991
Random Forest (500 árboles)	0,743	0,740	0,997
XGBoost	0,726	0,762	0,905

Modelos semisupervisados	Frequency-Based Seeds		
	Exactitud	Precisión	Recall
Naive Bayes	0,758	0,805	0,918
Kernel Linear SVM	0,771	0,808	0,932
Kernel Gaussian RBF SVM	0,793	0,794	0,997
Random Forest (100 árboles)	0,799	0,797	0,999
Random Forest (500 árboles)	0,801	0,799	0,999
XGBoost	0,794	0,819	0,948

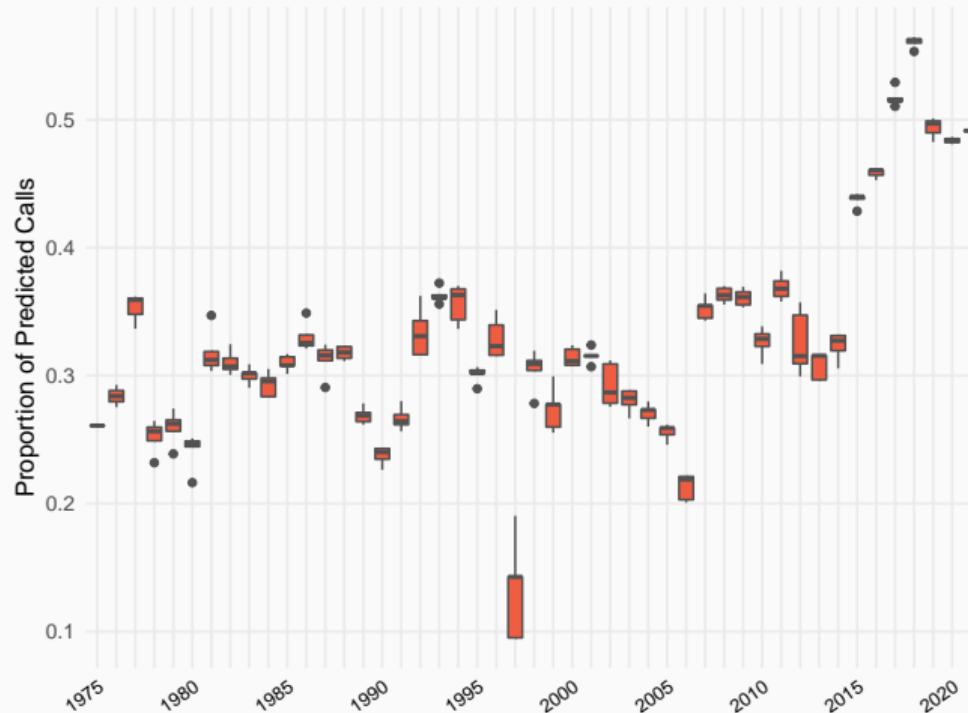
Rendimiento de los modelos semisupervisados (1975-2021)

Para la predicción utilizamos una estrategia similar a [Greene et al. \(2019\)](#) con una **rolling window** fija de cinco años.

Indicadores con 10-fold cross-validation en una ventana móvil de cinco años entre 1975 y 2021.



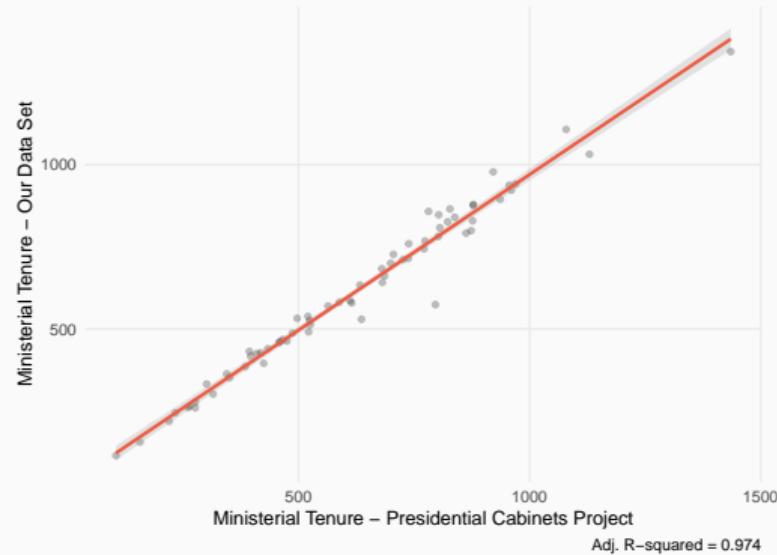
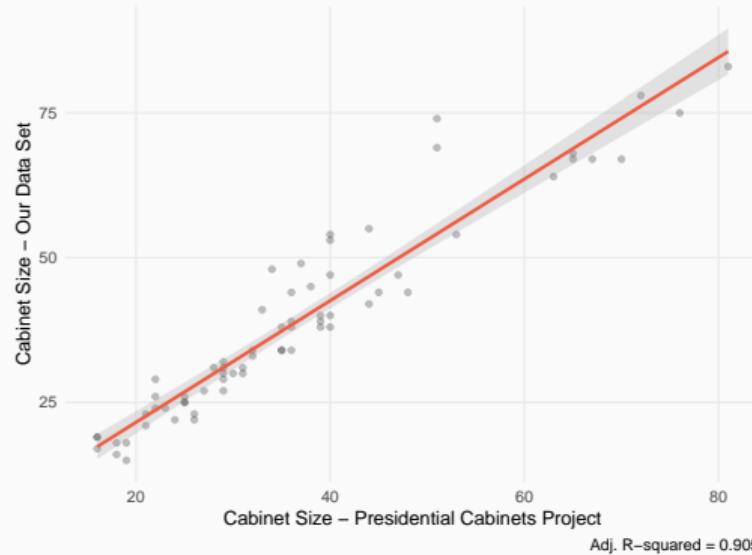
Predicción de peticiones a dimisiones ministeriales (1975-2021)



Proporción de predicciones sobre las menciones estimadas con un **algoritmo ensamblado semisupervisado Random Forest** (500 árboles) en la ventana móvil de cinco años.

Validación y aplicaciones

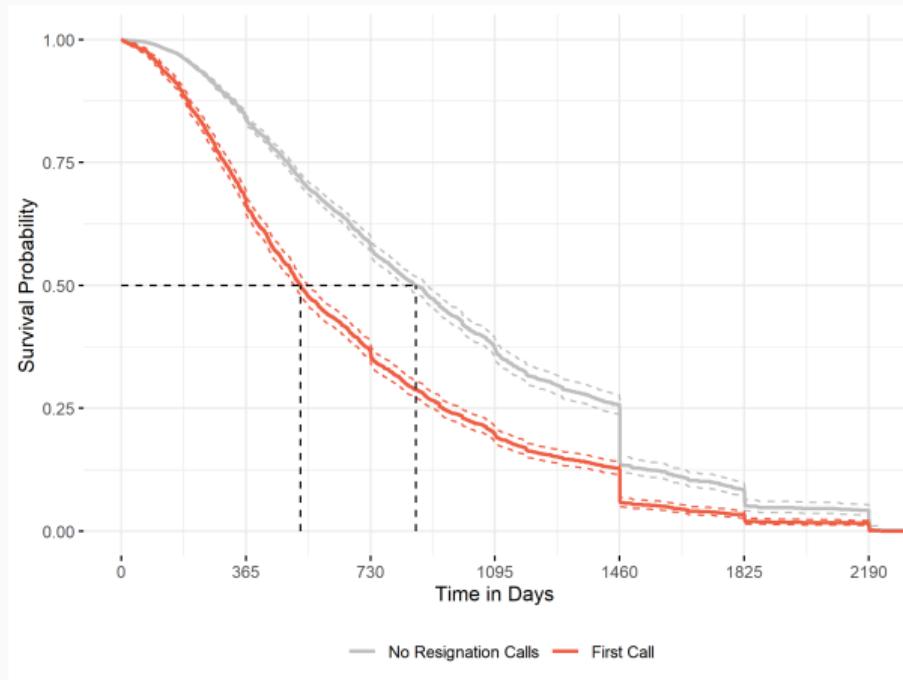
Comprobaciones de plausibilidad con conjuntos similares



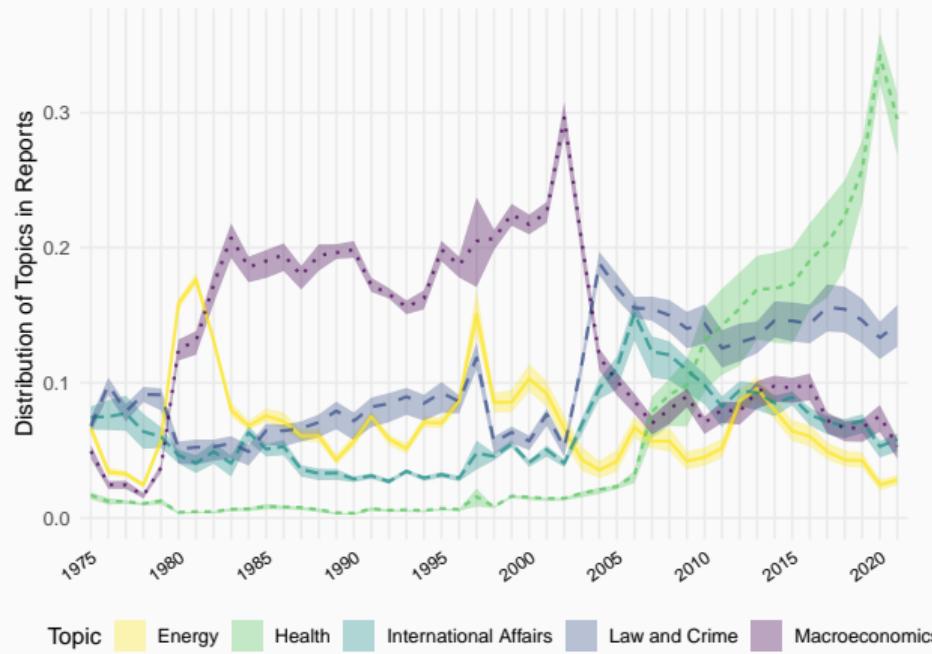
Nota: Mandatos concurrentes para comparación con el proyecto Presidential Cabinets de [Camerlo y Martínez-Gallardo \(2020\)](#).

Estimaciones de supervivencia

Podemos utilizar $F(t) = 1 - S(t) = \Pr(T \leq t)$ para obtener la incidencia de las salidas ministeriales durante T considerando $S(t)$ como una función de supervivencia.



Principales temas de la agenda en América Latina (1975-2021)



Hemos utilizado **Keyword Assisted Topic Modelling** (**keyATM**; véase Eshima et al., 2022). Consiste en especificar ciertas palabras clave para generar etiquetas antes del modelado de temas con un enfoque semisupervisado bayesiano.

Discusión

- Novedoso conjunto de datos con **indicadores totalmente inéditos**
- **Potenciales aplicaciones** más allá del estudio de coaliciones y rotación ministerial
- Aplicación pionera de **minería de datos y aprendizaje automático**
- Enfoque **semisupervisado** como alternativa para series de tiempo extensas

Agradecimientos y liberación de datos

Agradecimientos

Esta investigación ha sido financiada por la Agencia Nacional de Investigación y Desarrollo de Chile (ANID/PFCHA/72200340), el Fondo Muriel Wise del St Hilda's College de la Universidad de Oxford y el Premio de Apoyo a la Investigación Postdoctoral de la Sociedad de Estudios Latinoamericanos (Society for Latin American Studies, SLAS).

Mi agradecimiento a Petra Schleiter, Radoslaw Zubek y David Doyle por sus importantes comentarios y apoyo. También agradezco a Carla Cisternas y Diego Aguilar por su ayuda durante el proceso de revisión y validación de los archivos y al equipo de codificadores que realizaron el etiquetado para entrenar los modelos supervisados.

Liberación de datos

El conjunto de datos y sus metadatos estandarizados se depositarán en el repositorio digital  Oxford Research Archive for Data (ORA-Data). El conjunto de datos permanecerá en embargo hasta octubre de 2023 (por confirmar).

A partir de entonces, estará disponible para su reutilización bajo una licencia de acceso abierto  que permite compartir y adaptar el material sin restricciones adicionales, siempre y cuando se otorgue el reconocimiento correspondiente.

Referencias

Referencias I

- Berlinski, S., Dewan, T., y Dowding, K. (2010). The Impact of Individual and Collective Performance on Ministerial Tenure. *The Journal of Politics*, 72(2):559–571.
- Camerlo, M. y Martínez-Gallardo, C. (2020). The Presidential Cabinets Project. Research Project, Universidade de Lisboa and University of North Carolina at Chapel Hill. Available at <https://www.presidentialecabinets.org>.
- Camerlo, M. y Pérez-Liñán, A. (2015). Minister Turnover, Critical Events, and the Electoral Calendar in Presidential Democracies. *The Journal of Politics*, 77(3):608–619.
- Dewan, T. y Dowding, K. (2005). The Corrective Effect of Ministerial Resignations on Government Popularity. *American Journal of Political Science*, 49(1):46–56.
- Eshima, S., Sasaki, T., y Imai, K. (2022). Keyword Assisted Topic Models. Package, Harvard University and Massachusetts Institute of Technology. Package 'keyATM' version 0.4.1.
- Greene, K. T., Park, B., y Colaresi, M. (2019). Machine Learning Human Rights and Wrongs: How the Successes and Failures of Supervised Learning Algorithms Can Inform the Debate About Information Effects. *Political Analysis*, 27(2):223–230.

Referencias II

- Labelbox (2021). Labelbox. Online. Available at <https://labelbox.com>.
- Martínez-Gallardo, C. (2014). Designing Cabinet: Presidential Politics and Ministerial Instability. *Journal of Politics in Latin America*, 6(2):3–38.
- Sauvola, J. y Pietikäinen, M. (2000). Adaptive document image binarization. *Pattern Recognition*, 33(2):225–236.
- Singh, T. R., Roy, S., Singh, O. I., Sinam, T., y Singh, K. M. (2011). A New Local Adaptive Thresholding Technique in Binarization. *IJCSI International Journal of Computer Science Issues*, 8(6–2):271–277.
- Su, B., Lu, S., y Tan, C. L. (2010). Binarization of historical document images using the local maximum and minimum. En *DAS '10: Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, Boston. ACM Press.

Muchas gracias



Presentación compilada con **LATEX** y algunos ☕

↻ Descargar la última versión desde **GitHub**